INTELLIGENCE TESTS AND THE CLASSIFICATION OF PUPILS. II

F. S. BREED AND E. R. BRESLICH
University of Chicago

Two problems are discussed in this series of articles. In the preceding article an attempt was made to answer the question, How accurately do certain general-intelligence tests measure the intelligence of high-school pupils? In the present article the problem is, How accurately do these intelligence tests enable one to predict the educational achievement of high-school pupils?

Findings reported in the previous article.—In the first article the Otis, Terman, and Chicago group intelligence tests were found to have an inter-test correlation of .77. Thirty per cent of the pupils classified by one test were found to be out of place according to another. Eighteen per cent of the pupils classified by the tests were not classified according to their intelligence. In measuring an individual pupil, the tests differed, in average, 6 points on the Chicago scale, 11.1 points on the Otis scale, and 13.9 points on the Terman scale. The greatest difference between scores on two scales for the same pupil was 52.61 points. The Otis test gave more satisfactory results than the tests with which it was compared.

It is clear that high-school teachers must observe considerable caution in using the scores from these tests as measures of the intelligence of pupils. The data mentioned indicate the amount of error to expect. What caution, then, should be observed in inferring future educational achievement from the scores of these tests?

Two problems to be kept distinct.—In the consideration of this problem, comparisons will be presented between measurements made by means of intelligence tests at the beginning of a semester and measurements of educational achievement made during and at the end of the same semester. Two questions must be kept

distinct in the course of the discussion: (1) Is intelligence a satisfactory basis of classification? (2) Are scores from intelligence tests a satisfactory basis of classification? The answer to the first question must manifestly await further investigations. A satisfactory answer to the second question can be more easily obtained.

In order, however, that certain data relating to the first of these problems might have due consideration, it was decided to use a composite-intelligence score derived from the three intelligence tests in addition to scores from single tests. This was done on the principle that the average of several expert attempts to measure an object or reaction is probably more reliable than a single expert attempt. In other words, it is expected that the composite will

TABLE VI

METHOD OF TRANSMUTING SCALE SCORES INTO PERCENTILE SCORES

Number of Pupil	Score in Otis Test	Rank	Frequency Percentage	Sigma Score	Percentile Score
21..........	98	60	.02	.56	11.2
15..........	105	59	.03	.70	14.0
49..........	108	58	.05	.91	18.2
4..........	116	57	.07	1.07	21.4
47..........	118	55.5	.09	1.20	24.0

provide a more reliable measure of intelligence than any one of the three tests.

Computing a composite from the scores of the three tests.—Before the scores from the three tests can be averaged they must be expressed in the same unit. Several possible statistical methods of reducing the scores to a comparable basis were considered, two of which were selected and tested in regard to their reliability. These methods may be designated (1) transmutation of the three series of scores into units on a percentile scale and (2) transmutation of two series of scores into units on the third scale. For those who may be statistically interested, the steps of the transmutation process by Method 1 are illustrated in Table VI. In the first column of the table is shown the number of the pupil, in the second the Otis scores in the order of increasing magnitude, in the third the rank of the pupil, in the fourth the corresponding frequency percentage, in the fifth the corresponding score on a 5-sigma scale,

and in the sixth the score on a 100-point or percentile scale. A table of values of the probability integral was used to convert frequency percentages into corresponding values on the 5-sigma scale.[1]

Table VII shows for the same pupils the method of deriving the series of composite scores from the three series of percentile

TABLE VII

METHOD OF DERIVING THE COMPOSITE PERCENTILE INTELLIGENCE SCORE

Number of Pupil	Chicago Percentile Score	Otis Percentile Score	Terman Percentile Score	Composite Intelligence Score
21	21.4	11.2	20.0	17.5
15	28.0	14.0	20.0	20.7
49	24.0	18.2	11.2	17.8
4	11.2	21.4	14.0	15.5
47	36.8	24.0	28.0	29.6

scores. The composite-intelligence score for any pupil was obtained by averaging the three percentile scores for that pupil. In Table VII pupil 21 is shown to have percentile scores of 21.4, 11.2, and 20, in the Chicago, Otis, and Terman group tests, respectively, from which his composite percentile intelligence score of 17.5 is derived.

TABLE VIII

RELIABILITY OF DERIVED PERCENTILE SCORES

Series of Scores	Coefficient of Correlation	P. E.
Chicago and its derived percentile	.996	.0007
Otis and its derived percentile	.990	.0017
Terman and its derived percentile	.979	.0036

As a test of the reliability of this transmutation process, the correlation was computed between each original series of scores and the percentile series derived therefrom. The coefficients and probable errors appear in Table VIII. The coefficients are all extremely high and the probable errors negligible. This indicates that in each case the derivative score is an accurate substitute for

[1] See H. O. Rugg, *Statistical Methods Applied to Education*, p. 392. Boston: Houghton Mifflin Co.

the original from which it was derived. And if the three scores of each individual are thus accurately represented, the composite derived therefrom should accurately represent the individual's average score in the three tests.

The reliability of this series of composite scores was further checked by comparison with a series of scores derived by the second method of compositing mentioned. The Chicago and Terman scores of each pupil were transmuted into Otis units, and these two derived scores were averaged with the pupil's actual Otis score. The transmutation was performed by the use of the regression equation, $y = r \frac{\sigma y}{\sigma x} x$, in which y represents a pupil's

TABLE IX

METHOD OF TRANSMUTING CHICAGO AND TERMAN SCORES INTO OTIS VALUES

Number of Pupil	Actual Otis Deviation	Transmuted Chicago Deviation	Transmuted Terman Deviation	Deviation of Composite Score
1	− 6.0	−22.39	+ 2.03	− 8.79
2	+ 6.0	+ 4.44	+13.73	+ 8.06
3	+34.0	+ 8.14	+16.85	+19.66
4	−35.0	−42.74	−46.33	−41.36
5	+17.0	+42.37	+26.21	+28.53

deviation from the mean of one series of scores, x his deviation from the mean of another series, σy the standard deviation of the first series, σx the standard deviation of the second series, and r, which in this case is given the value of unity, the coefficient of correlation between the two series. After the deviation of each individual score from the median of its distribution was found—that is, the y and x values in the formula—it became a simple matter to compute for any pupil's deviation in one test the corresponding deviation, under conditions of perfect correlation, in another. In other words, it was thus possible to compute for any score in one test the equivalent score in either of the other tests.

Table IX illustrates the results of these computations. In the first column is given the number of the pupil, in the second his deviation from the median of the Otis series, in the third his Chicago deviation transmuted into Otis units, in the fourth his

Terman deviation transmuted into Otis units, and lastly the average of these three deviations. The average, for pupil No. 1, −8.79, indicates that this pupil's composite intelligence score is 8.79 points below the median score (151) of the group in the Otis test.

This series of composite values was used to check the reliability of the percentile composite series. Low correlation between the two series would indicate that at least one of the two methods of compositing is unreliable. High correlation between the two would justify the conclusion that both methods are reliable. The coefficient of correlation was found to be .998, with a negligible probable error, indicating the satisfactory reliability of both methods. The percentile composite scores were now used to represent the average measurement of the three tests in the later study of the relation between scores on intelligence tests and measures of educational achievement. This method of compositing was also used later to derive an arithmetical-ability score from results secured in tests of computational skill and arithmetical reasoning, and to combine scores in arithmetical ability and general intelligence.

The relation between standing in general-intelligence tests and achievement in school tests.—To the fifty-four seventh-grade pupils and the sixty ninth-grade pupils, who had taken the three intelligence tests near the beginning of the semester, certain school tests were administered monthly during the semester and a final examination at the end of the semester. These tests, like all other tests of educational achievement used in this study, were limited to the field of mathematics. The school tests referred to here were prepared by Mr. Breslich. They were alike for a given phase of the work in all sections of the same grade, were administered under the direction of the author, and were scored by a key prepared by him to accompany them.

SAMPLE TEST WITH KEY FOR SCORING

I. A building 45 ft. high casts a shadow 55 ft. long. Find the angle of elevation of the sun by means of a scale drawing.

Select the scale	1
Draw the figure	1
Find the angle	1

II. Find by similar triangles the height of a building which casts a shadow 43 ft. long when a vertical seven-foot pole casts a shadow 9 ft. long.

Make the sketch 1
Write proportion 1
Solve the equation 1

III. Solve

$$\frac{x}{10}=\frac{19}{6}$$

Multiply by 10 1
Reduce 1

Solve

$$\frac{2r}{3}+\frac{3r}{4}=23$$

Multiply by 12 1
Reduce 1
Collect terms 1
Divide 1

IV. Find the angle of elevation by means of the tangent ratio when a building 118 ft. high casts a shadow 143 ft. long.

$\tan x=\frac{118}{143}$ 1
$\tan x=.81$ 1
$x=39^\circ$ 1

The results from the school tests were employed in two ways. First, the average score of each pupil was found in the monthly tests only. Second, the average score was found for the monthly test and the final examination, each monthly test being given a weight of one and the examination a weight of two. The correlation between the percentile composite intelligence scores and each school-test series was then computed for both grades, with the results shown in Table X. It is observed that all the coefficients

TABLE X

Correlation between School Tests and Percentile Composite Intelligence Scores

Series of Scores	Seventh Grade	Ninth Grade
Monthly tests and percentile composite	.365	.327
School tests and percentile composite	.391	.315

are between .30 and .40, indicating no close relationship between the intelligence scores of the pupils, as represented by the percentile composite, and achievement in the school tests. If these school-test results are accepted as valid measures of achievement, the composite-intelligence scores clearly do not constitute an accurate basis of classification.

The degree to which classifications of pupils based upon the two series of scores are at variance with each other is shown in Figures 6 and 7. Each pupil is indicated by a small square in the diagram. The upper number in each square is the pupil's intelligence score, the lower his school-test score. The two heavy lines running upward through the figure mark the divisions between sections. The shaded squares show the displaced pupils, that is, pupils who would be in other sections if their classification were based on their record in the school tests.

In Figure 6, representing the ninth-grade group, the pupils are classified by the intelligence composite into three sections, corresponding to the three sections in ninth-grade mathematics. Out of the total of fifty-one pupils with complete records in both series of tests, twenty-eight, or 55 per cent, are found displaced. Figure 7 in similar manner shows the amount of pupil displacement in the seventh grade when forty-six pupils are classified according to the composite-intelligence scores into two sections. Eighteen, or 39 per cent, are found displaced. It should be noted that Figure 6 represents the situation where the coefficient of correlation was .31 and Figure 7 where it was .39, these being the lowest and highest coefficients respectively in Table X.

In attempting to draw practical conclusions from the foregoing data, consideration must be given not only to the reliability of the intelligence scores, but to the reliability of the school-test scores as well. In investigations of this sort scores from conventional school tests and the ordinary marks of teachers are commonly accepted as representative of a pupil's achievement in a subject. The same sort of critical caution should be observed with reference to measures of educational achievement that is now observed in dealing with measures of intelligence.

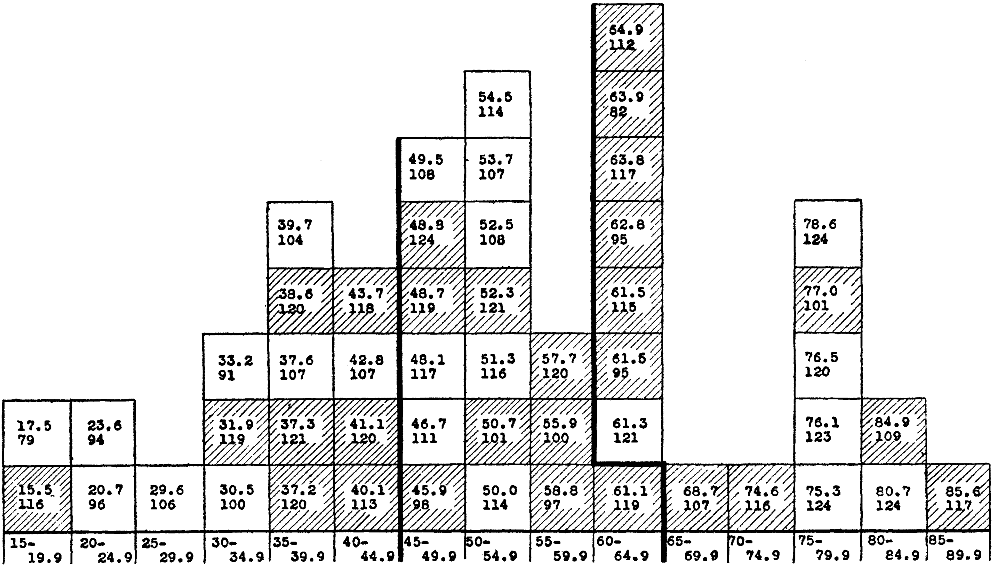

FIG. 6.—Displacements of ninth-grade pupils classified into three sections according to composite-intelligence scores. School tests were used as the criterion of educational achievement. Hatched squares represent displaced pupils. Coefficient of correlation, .31; percentage of displacement, 55.

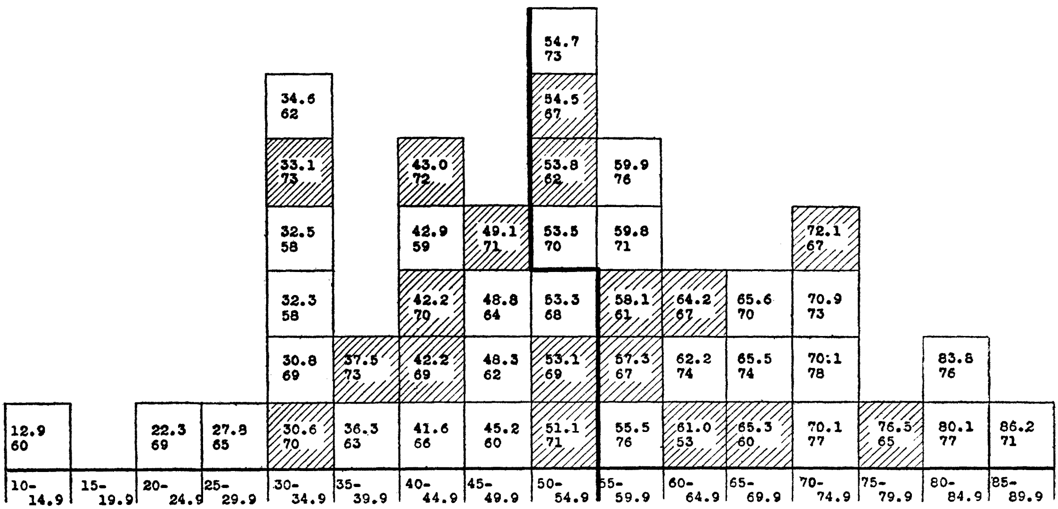

FIG. 7.—Displacements of seventh-grade pupils classified into two sections according to composite-intelligence scores. School tests were used as the criterion of educational achievement. Hatched squares represent displaced pupils. Coefficient of correlation, .39; percentage of displacement, 39.

The school tests administered in this investigation must be considered in connection with the particular system of instruction of which they form a part. In the University High School there is unusually strong emphasis on the mastery of essentials. The tests are constructed to determine whether the pupils have mastered certain definitely outlined units of work. The time allowance on a test varies for different pupils; that is, each is given as much time as he needs, within reasonable limits, to do his best on a unit. Thus achievement tends to be equalized, individual differences to be minimized, and ability to be measured apart from the factor of time. A considerable percentage of pupils commonly obtain an identical maximum score. While it is very probable that these tests serve admirably the purpose for which they are intended, it is quite clear that they are designed to meet a special situation, one that cannot be regarded as typical in public high schools. It seemed, therefore, expedient to make a further study of the educational achievement of one of these groups of pupils. The ninth grade was selected for this purpose.

At the beginning of this investigation it was thought that careful measurements of the previous achievement of these pupils in mathematics might possibly constitute a more satisfactory basis of classification for further work in mathematics than the scores of intelligence tests. Accordingly, along with the intelligence tests given at the beginning of the semester, the Cleveland Survey Arithmetic Tests were administered to determine the computational skill of the pupils. The arithmetical-reasoning tests administered as parts of the Otis and Terman group tests of intelligence were also used to measure ability in arithmetical reasoning. In scoring the Cleveland tests, the weights given by Counts[1] for the problems of the fifteen different component tests were used. The average score in the two reasoning tests was taken to represent arithmetical-reasoning ability. A score in *arithmetical ability* was obtained by averaging the scores in computation and reasoning with equal weight after each series had been transmuted, as previously described, into units on a percentile scale.

[1] George S. Counts, *Arithmetic Tests and Studies in the Psychology of Arithmetic*, p. 28. "Supplementary Educational Monographs," Vol. I, No. 4. Chicago: University of Chicago Press, 1917.

At the end of a semester a third check on mathematical achievement was secured. A test was given which was composed of the following examples taken from the Hotz first-year algebra scales,[1] Series A: addition and subtraction: 1, 5, 8, 13—5 minutes; multiplication and division: 1, 2, 3, 7, 9, 11, 16—14 minutes; equation and formula: 1, 3, 4, 6, 11—6 minutes; problems: 1, 2, 4, 7—7 minutes; graphs: all—25 minutes. Examples were selected which belonged within the field covered by the pupils during the semester. The test was uniformly administered to the three sections by Mr. Breslich; the papers were scored by a trained assistant, and each example in the scoring was given the value determined for it by Hotz.

Finally, inasmuch as industry seems to be one of the most important factors in scholarship, careful ratings of industry were secured from the teachers of these ninth-grade sections in mathematics. Blanks were provided with the following directions:

> Rate each of the pupils in each of the following characteristics, using the numbers 1, 2, 3, 4, and 5; 1 being the lowest rating and 5 the highest: (*a*) promptness in beginning new work, (*b*) concentration on the work once begun, (*c*) perseverance in doing assigned tasks, (*d*) accomplishment of more than the minimum requirement, (*e*) attention to questions raised and suggestions made during the class period.

Comparison of school tests and Hotz examples.—In Table XI results from the school tests and Hotz examples are compared in regard to the closeness of their relationship with other factors ordinarily considered as important for achievement in ninth-grade mathematics. The school tests in this case include both the monthly tests and the final examination. It will be seen that the Hotz examples yield a measurement more closely related to the intelligence composite than do the school tests. The respective coefficients are .56 and .31. Both tests give results fairly closely related to industry, with no appreciable difference in the closeness of the relationship. Each shows a closer relation to intelligence and industry combined than to either one of these factors alone. Industry in this computation was given a weight of four as against one for intelligence. There is no significant difference in the size

[1] H. G. Hotz, *First-Year Algebra Scales,* pp. 5 f. "Teachers College Contributions to Education," No. 90. New York: Teachers College, Columbia University, 1918

of these coefficients. In relation to arithmetical ability, however, there is a marked difference between the tests. For the Hotz examples there is a coefficient of .43, and for the school tests a coefficient of .18. In the last item of the table is presented the relationship between the results from each of these tests and an average of the scores representing arithmetical ability and general intelligence. The last two series of scores were expressed in a common unit by reduction to a percentile basis before being averaged. The correlation in this case is much higher for the Hotz examples than for the school tests.

In summarizing the differences between the two tests, it may be noted that the Hotz test (1) is designed to meet a situation more commonly found in typical high schools, (2) equalizes the time factor in achievement, (3) distributes the pupils more widely in the upper range of the scale, (4) yields results more closely related to arithmetical-ability scores, and (5) yields results more closely related to scores from intelligence tests.

TABLE XI

School Tests and Hotz Examples in Relation to Various Factors in Educational Achievement

Factors	Correlation with School Tests	P. E.	Correlation with Hotz Examples	P. E.
Intelligence composite	.31	.08	.56	.06
Industry rating	.53	....	.55	....
Intelligence composite and industry rating × 4 (one section)	.61	.09	.62	.09
Arithmetical ability	.18	.09	.43	.08
Arithmetical ability and intelligence composite	.29	.09	.58	.06

The last two differences are probably due in large part to the first three. Lack of differentiation among the best pupils in the school tests would seriously affect the correlation with arithmetical-ability and general-intelligence scores. On account of this fact it does not seem profitable to enter into a further discussion of the significance of the difference between the correlation coefficients for the two tests. The main differences between the tests are clear and seem justified, for the most part, by the different functions

which they are designed to serve. The fundamental point seems to be that a test designed to discriminate mastery in the manner of the school tests is not suitably designed to discriminate the contribution to that mastery made by arithmetical ability and intelligence. In view of these considerations the Hotz test was employed as a second criterion of educational achievement.

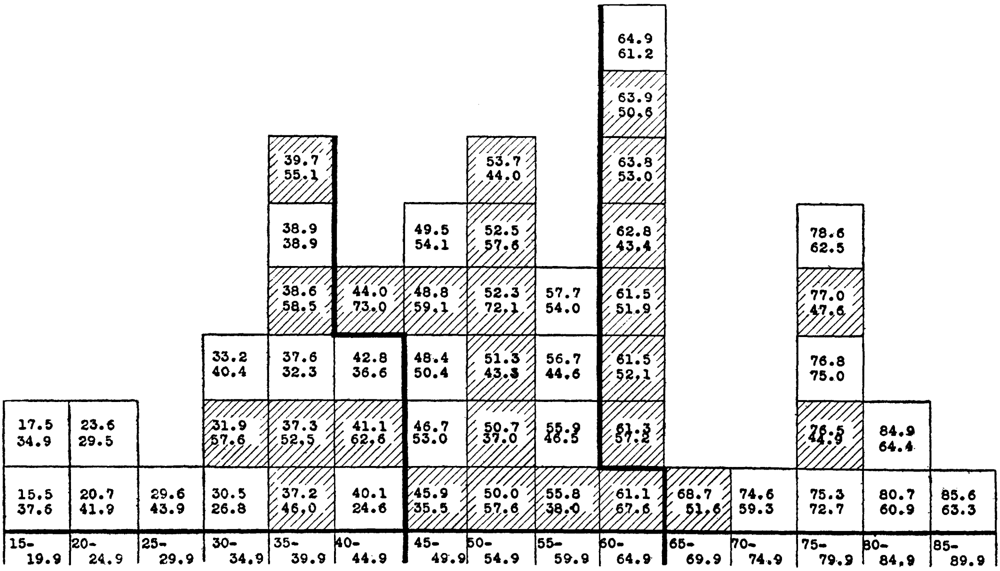

FIG. 8.—Displacements of ninth-grade pupils classified into three sections according to composite-intelligence scores. Hotz examples were used as the criterion of educational achievement. Hatched squares represent displaced pupils. Coefficient of correlation, .56; percentage of displacement, 51.

A comparison of various bases of classification.—The last coefficients in Table XI deserve special comment. These values have an interesting bearing on the problem of classifying ninth-grade pupils for work in mathematics. They may be considered in connection with the larger question, Which of the following, according to the foregoing data, would constitute the most satisfactory basis of classifying these pupils? (1) Intelligence tests, (2) arithmetical-ability tests, (3) intelligence and arithmetical-ability tests combined. Attention, of course, is purposely confined to bases available for use at the beginning of high-school work.

It has already been observed that the measures of arithmetical ability did not constitute a very satisfactory index of the later success of these pupils in mathematics. Further, combining the arithmetical-ability scores with the intelligence scores did not provide an appreciably more satisfactory basis of classification than the intelligence scores alone. The last two modes of classification are illustrated graphically in Figures 8 and 9. Figure 8 represents the ninth-grade pupils classified by the intelligence composite, with displacements checked against a classification according to the Hotz test. The percentage of displaced pupils was 51. When the same pupils were classified according to the combined scores in intelligence and arithmetical ability, the percentage of displaced pupils, Hotz scores being used as the criterion, was again 51. This is shown in Figure 9. These data on pupil displacement confirm the conclusion reached, namely, that a classification of the pupils based on intelligence scores would be quite as satisfactory as a classification based on a combination of intelligence and arithmetical-ability scores.

In order that the intelligence-composite and the Hotz-examples series of scores may be seen more clearly in their relation to other important measures, Table XII is provided. The correlation with

TABLE XII

INTELLIGENCE COMPOSITE AND HOTZ EXAMPLES IN RELATION TO OTHER FACTORS IN THE NINTH GRADE

Factor	Correlation with Intelligence Composite	P. E.	Correlation with Hotz Examples	P. E.
Arithmetical ability	.51	.058	.43	.077
Industry rating	.57	....	.55	....
School tests	.31	.081	.58	.066

industry is about the same for both, the intelligence composite correlates more highly with arithmetical ability and less highly with school tests. The coefficient of correlation with each other is .56.

While the intelligence composite was found to be as satisfactory a basis of classification at the beginning of ninth-grade mathematics as any of the other bases tested, it must be noted that most high

schools would find the administering of three intelligence tests and the computation of composite scores a rather tedious process. It would be interesting to know what the loss in accuracy of classification would be, if any, in case the Otis test were used instead of the intelligence composite. The Otis test was therefore substituted. The correlation between the Otis and Hotz scores was .53. The pupil displacement was 51 per cent, indicating no loss in accuracy. This is graphically shown in Figure 10.

To summarize the portion of the discussion dealing with the basis of classification, Table XIII is presented. It is seen that there

TABLE XIII

SUMMARY OF DATA ON CORRELATION AND PUPIL DISPLACEMENT

	Coefficient of Correlation	Percentage of Displacement
Otis group and Hotz examples	.53	51
Intelligence composite and Hotz examples	.56	51
Intelligence composite and arithmetical ability; and Hotz examples	.58	51

is in the correlation coefficients a suggestion of improvement in the base as one proceeds from the first to the third of these comparisons. The figures on displacement, however, furnish no evidence with regard to relative superiority. The Otis test actually classified the pupils as well as either of the other means employed.

A displacement of 51 per cent in the foregoing case seems large. One may be inclined to think that measures of intelligence should forecast scholarship with greater accuracy. They probably would if intelligence and scholarship were accurately measured. Some psychologists, however, seem to believe that with perfect instruction and perfect measurement the pupil displacement in a case of this kind would be reduced to zero. On the basis of this view it has even been suggested that teachers be rated as efficient or inefficient according to the discrepancy between the intelligence and educational-achievement measurements of their pupils. The results indicate that there is no reasonable prospect of the success of such a scheme. Clearly, scholarship is not a matter of intelligence alone. It is a product, as well, of such powerful emotional factors as interest and such volitional factors as perseverance. These emotional and volitional elements, only slightly if at all measured

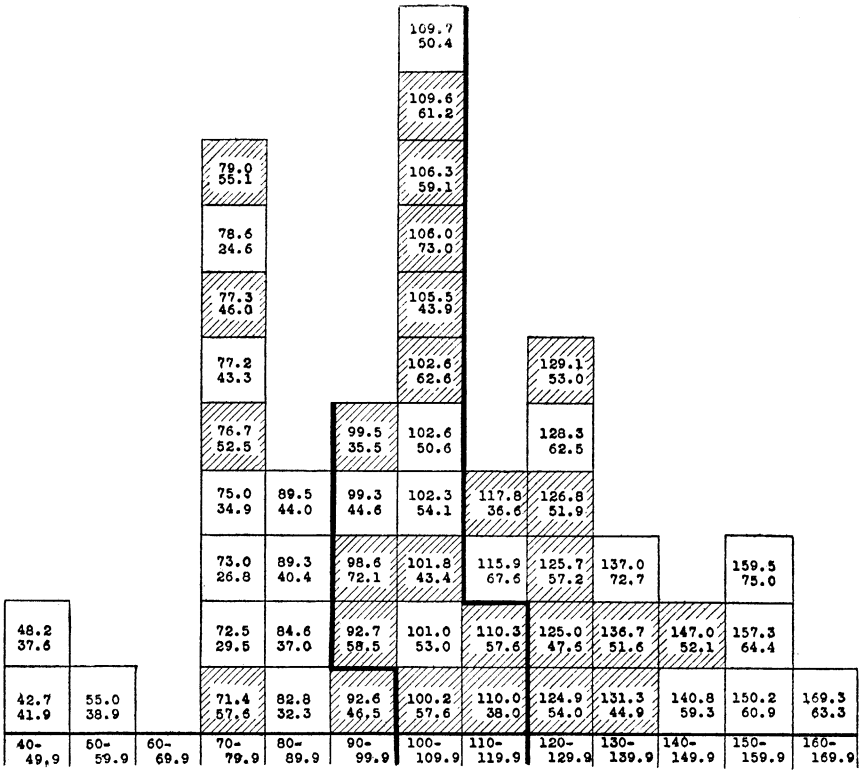

FIG. 9.—Displacements of ninth-grade pupils classified into three sections according to the combined scores in intelligence and arithmetical ability. Hotz examples were used as the criterion of educational achievement. Hatched squares represent displaced pupils. Coefficient of correlation, .58; percentage of displacement, 51.

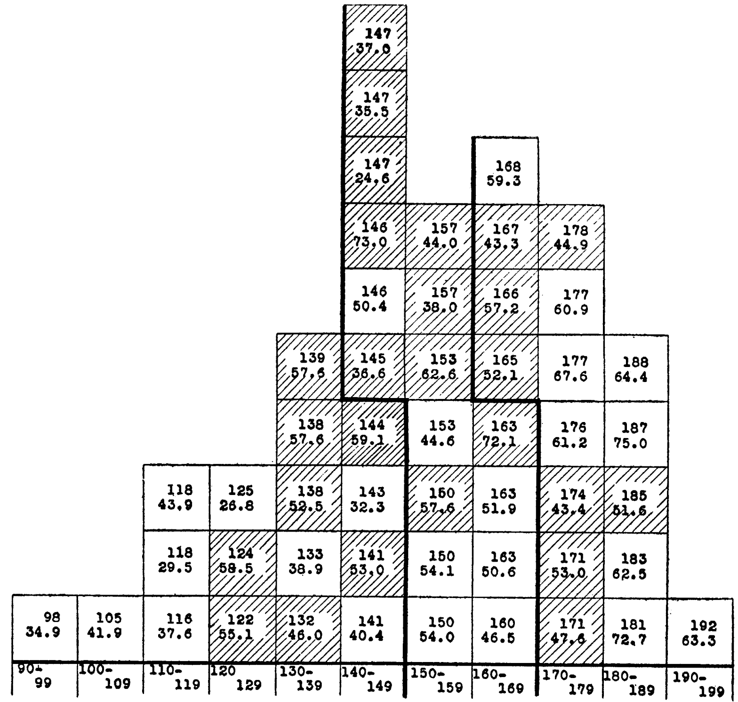

FIG. 10.—Displacements of ninth-grade pupils classified into three sections according to scores in the Otis group intelligence test. Hotz examples were used as the criterion of educational achievement. Hatched squares represent displaced pupils. Coefficient of correlation, .53; percentage of displacement, 51.

by intelligence tests, remain outside the field of intellect, to hamper or quicken the progress of the pupil in any subject. Not only does the best psychological theory seem to demand that these two aspects of our mental life be distinguished from the functions of intelligence, but it seems further to support the view that a gift in one does not necessarily imply a gift in all. Will and emotions, apparently, can no more be produced by facile instruction than can intellect.

The problem of classifying pupils by intelligence tests is obviously, then, complicated by the following conditions: (1) imperfect instruments for measuring intelligence, (2) imperfect instruments for measuring educational achievement, (3) imperfect correlation between intelligence and interest, (4) imperfect correlation between intelligence and will, (5) imperfect stability of the pupil, (6) imperfect instruction.

It seems probable, therefore, that the most accurate measurements of intelligence will not provide a reliable basis for classification under the most perfect school conditions. We have found that the Otis test failed to classify at least 13 per cent of a group according to their intelligence. It failed to classify 51 per cent according to their educational achievement. The inaccuracy in the measurement of intelligence does not account for all of the error in the second case. If the scholarship test in the second instance be assumed to have an error as large as that of the Otis test, and it is not conceded that it is larger, the disparity between intelligence and scholarship is not yet explained. There is good reason from these data to believe that other factors such as those enumerated are involved in the situation and make the problem a vastly more complex one than positing a perfect relationship between two abilities such as intelligence and scholarship and measuring one of them.

All one should expect from the group tests of intelligence, so far as the general problem under discussion is concerned, is that they provide a preliminary classification, which will be subject to rectification as the scholarly ability of the pupils becomes known. This they did in the present study more economically than any other means tried, and otherwise as satisfactorily. Other things being equal, the accuracy of such classifications will probably

increase considerably as the reliability of the measuring instruments, both psychological and educational, is increased.

SUMMARY AND CONCLUSIONS

1. This article is devoted to a study of the problem, How reliable are intelligence tests as a basis for classifying high-school Freshmen in mathematics?

2. Measures of intelligence are represented by the scores from the most reliable of three intelligence tests and by a composite of the three tests. The composite was assumed to provide a more reliable measure of intelligence than any single test.

3. A series of school tests and a test composed of examples from the Hotz algebra scales were used as the principal criteria of educational achievement.

4. Other data employed were industry ratings and a composite representing arithmetical ability.

5. The average correlation between the composite-intelligence scores and the school tests was .35. Pupil displacement for the ninth grade, when divided into three classes, was 55 per cent; for the seventh grade, when divided into two classes, 39 per cent.

6. The correlation between the composite-intelligence scores and the Hotz test was .56. Pupil displacement for the ninth grade, when divided into three classes, was 51 per cent.

7. The Otis test classified the pupils more satisfactorily than arithmetical ability scores, and as satisfactorily as either the intelligence composite or a combination of the intelligence-composite and arithmetical-ability scores.

8. Neither the composite intelligence scores nor the best of the intelligence tests provided a reliable basis for *permanent* classification. The error was in no case less than 50 per cent for a three-sectional classification in the ninth grade.

9. When the Hotz examples were used as the criterion of educational achievement, the Otis test provided a basis as satisfactory for a *temporary* classification as any other test or combination of tests tried, and did this more economically.

10. Intelligence is only one of a number of important factors in educational achievement.